5.1 Analysis of Terminal Digits in Underlying Data for Wang *et al*. 2012 *J. Neurosci* paper.

Paul S. Brookes, University of Rochester Medical Center April 7th 2023

This document (and its accompanying Microsoft Excel spreadsheet, "*TDA_JNeurosci.xlsx*") describes the application of "terminal digit analysis" to the underlying data sets used to plot graphs in the Figures of Wang *et al.* (2012) *J. Neurosci.* **32:** 9773-9784. Note that in several recovered files, "C105" refers to PTI-125/simufilam.

Methodology

A series of Microsoft PowerPoint files were identified that contain the original versions of charts and graphs from the paper figures. In the .PPT files, the charts also contain the embedded spreadsheet data from Microsoft Excel (.XLS), so it is possible to extract the underlying data for analysis.

Terminal digit analysis (TDA) is a tool to determine the probability that a data set arose naturally via a stochastic process [Mossimann *et al.* (2002) *Accountability in Research.* 9: 75-92. DOI: 10.1080/08989620290009530]. TDA has been used to uncover data manipulation, most famously in the case of Dr. Brian Wansink, a nutritional scientist from Cornell University who had 18 papers retracted for fabricating data.

In principle, for truly random data the last digit on the right-hand side of each number should be evenly distributed 0 – 9. For example, if a data set contains 150 numbers, we should expect to find each of the digits 0 – 9 approximately 15 times in the right-most or terminal position. TDA exploits the fact that humans are not good at generating truly random numbers (e.g., when entering random numbers on a computer keyboard, muscle-memory or keyboard layout may cause certain digits to be under- or over-represented in the results).

A Chi-Squared test is performed to compare the actual vs. expected distribution of terminal digits. The result is presented as a p-value, which is the probability that the observed distribution could have occurred naturally (by chance). The smaller this number, the less likely the distribution is natural and the more likely the data were fabricated by a human.

In TDA the number zero is sometimes ignored. This is because the method used to extract the right-most digit in Microsoft Excel (=RIGHT(text, [num_chars])) does not count zeros in numbers with decimal places, and the data set here includes a mix of items with and without decimal points. For example, the values 538.0 and 538.1 have terminal digits of 0 and 1 respectively, but Excel handles the first value terminal digit as an 8. However, for the values 538 and 540, Excel would return 8 and 0 respectively, because the zero in 540 is a terminal digit before the decimal point. Removing zeros from TDA somewhat weakens its statistical power (because now we are only looking at the distribution of digits 1 – 9 instead of 0 – 9), but it also means any findings on the remaining digits 1 – 9 are likely to be even more meaningful.

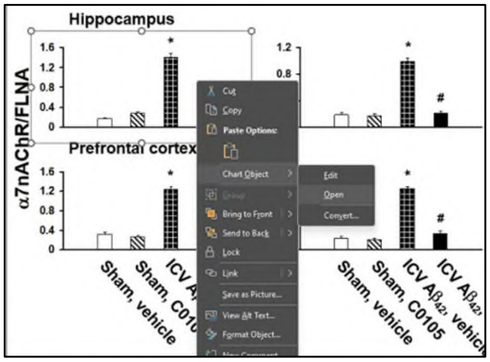
The .PPT files

The following PowerPoint files contain the original data for the figures in the paper:

Fig. 1A/B	*ICV-Ab42-FLNAa7TLR4-hp-pfc.ppt* (modified 2011-04-17, 21:47)
Fig. 1C/D	*ICV-ptau-c0105.ppt* (modified 2011-05-18, 07:36)
Fig. 2A/B	*FLNA-a7TLR4-CAD11.ppt* (modified 2011-04-17, 11:51)

Fig. 3A/B	*ICV-Ab42-PTI125-Ca influx.ppt* (modified 2011-05-14, 20:21)
Fig. 3C/D	*Ca-influx-C-AD 11 pairs.pp*t (modified 2011-05-14, 20:36)
Fig 5	*NMDAR signaling-C-AD 11 pairs quan.ppt* (modified 2011-05-16, 18:25)
Fig 6A	*IR signaling-ICVAb42 C0105.ppt* (modified 2011-06-09, 11:55)
Fig 6B	*IR signaling-C-AD 11 pairs quan-blots-final.ppt* (modified 2011-08-30, 07:11)
Fig 7	*cytokine-icv Ab-C0105.ppt* (modified 2011-06-09, 11:59)
Fig 9A	*Ab42-a7 complexes-CAD11-with mw marker.ppt* (modified 2011-03-28, 18:41)
Fig 10	*FLNA-a7-TLR4-pentapeptide control w quant-PTI125.ppt* (modified 2011-09-20, 08:56)
Fig 11	*FLNA-pTau-pentapeptide control.ppt* (modified 2011-06-16, 08:52)
Fig 12	*lymphocyte data.ppt* (modified 2011-10-14, 08:10)

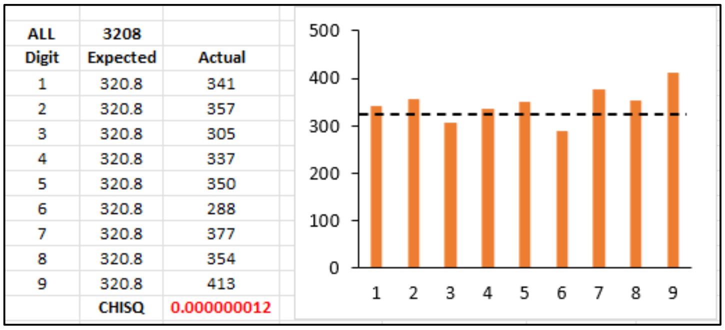
A key feature of the PowerPoint files is they contain charts that have been pasted in directly from Microsoft Excel spreadsheets, thus retaining the original data used to plot the charts. As shown here →, by right-clicking on the chart it is possible to extract the underlying .XLS file with its data.



Terminal Digit Analysis (TDA)

From the .PPT files, a total of 13 .XLS spreadsheets were extracted, containing a total of 3519 original data points. Data for Figure 7 were excluded from this analysis because the sheet contained data expressed as decimals rounded to the nearest half (0.5, 1.0, 1.5, etc.) and so the number 5 was over-represented as terminal digit (31% of all numbers).

The remaining figures contained 3208 data points, so each digit 0 – 9 would be expected to occur in the right-most position 320.8 times. The accompanying file "*TDA_JNeurosci.xlsx*" file contains the terminal digit analysis on the data set as a whole, and on the data for each individual Figure. On the right → is shown the expected vs. actual distribution for whole set. The dotted line is the expected number of times each digit occurs, and the orange bars the actual distribution. The data show that the number 9 occurs about 29% more frequently than expected, while the number 6 occurs at only 89% of its expected frequency.



ALL	3208	
Digit	Expected	Actual
1	320.8	341
2	320.8	357
3	320.8	305
4	320.8	337
5	320.8	350
6	320.8	288
7	320.8	377
8	320.8	354
9	320.8	413
	CHISQ	0.000000012

The CHISQ test at the bottom of the table returned a p-value of 0.000000012. Statisticians typically consider any p-value lower than 0.05 (meaning a 5% probability) to be *statistically significant*. In other words, a 1 in 20 chance the result is not real. In this case, the probability of the number distribution having occurred by chance as expected for experimental data, is 1 in 100 million. This is highly suggestive that the data set has been fabricated.

Breaking down the analysis to individual figures →, statistically significant findings were also seen for the data sets underlying Figure 2A/B, 3C/D, 5, 6A, 6B and 9A. In other words, it is unlikely these distributions came about from random or stochastic processes.

A notable pattern across all the data is an over-representation of the numbers 7 and 9.

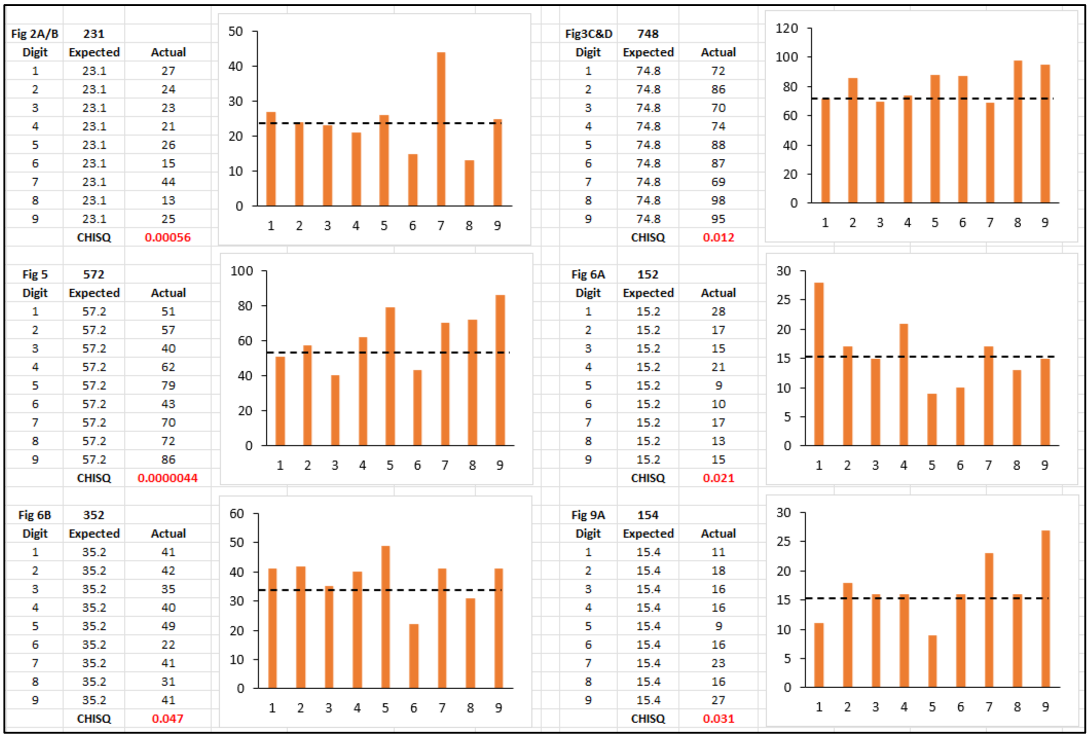


Fig 2A/B	231	
Digit	Expected	Actual
1	23.1	27
2	23.1	24
3	23.1	23
4	23.1	21
5	23.1	26
6	23.1	15
7	23.1	44
8	23.1	13
9	23.1	25
	CHISQ	0.00056

Fig3C&D	748	
Digit	Expected	Actual
1	74.8	72
2	74.8	86
3	74.8	70
4	74.8	74
5	74.8	88
6	74.8	87
7	74.8	69
8	74.8	98
9	74.8	95
	CHISQ	0.012

Fig 5	572	
Digit	Expected	Actual
1	57.2	51
2	57.2	57
3	57.2	40
4	57.2	62
5	57.2	79
6	57.2	43
7	57.2	70
8	57.2	72
9	57.2	86
	CHISQ	0.0000044

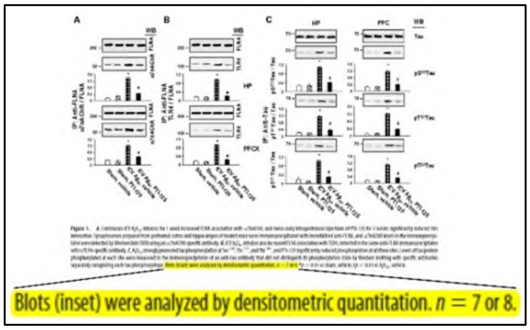
Fig 6A	152	
Digit	Expected	Actual
1	15.2	28
2	15.2	17
3	15.2	15
4	15.2	21
5	15.2	9
6	15.2	10
7	15.2	17
8	15.2	13
9	15.2	15
	CHISQ	0.021

Fig 6B	352	
Digit	Expected	Actual
1	35.2	41
2	35.2	42
3	35.2	35
4	35.2	40
5	35.2	49
6	35.2	22
7	35.2	41
8	35.2	31
9	35.2	41
	CHISQ	0.047

Fig 9A	154	
Digit	Expected	Actual
1	15.4	11
2	15.4	18
3	15.4	16
4	15.4	16
5	15.4	9
6	15.4	16
7	15.4	23
8	15.4	16
9	15.4	27
	CHISQ	0.031

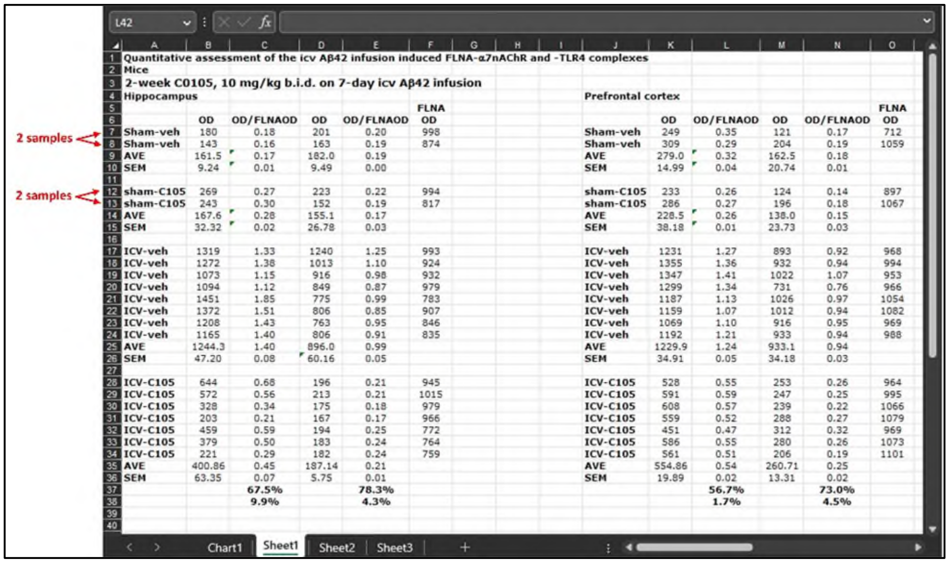
Additional Evidence Suggestive of Fabrication

The legend for Figure 1 on the paper → states that the number of experimental replicates (N) was 7–8. This implies that for all groups examined, there were between 7 and 8 independent biological samples.



Blots (inset) were analyzed by densitometric quantitation. *n* = 7 or 8.

The file *"ICV-Ab42-FLNa7TLR4-hp-pfc.ppt"* (modified 2011-04-17 21:47) contains the original western blot images for this Figure, as well as original graphs/charts presented below each blot. The .XLS file within the .PPT (shown below) contains data values, with each row corresponding to a single sample. Although the 3rd and 4th groups down the page (ICV vehicle and ICV C105, rows 17-36) contain the expected 8 and 7 samples respectively, the sham control groups at the top of the sheet only contain 2 samples each (indicated in red). As such, the number of replicates is far lower than claimed in the Figure legend.

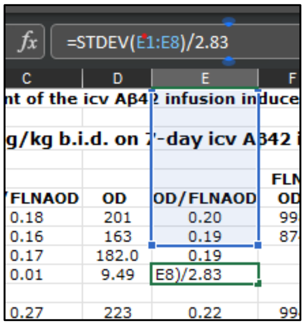


2 samples

2 samples

	A	B	C	D	E	F	J	K	L	M	N	O
1	Quantitative assessment of the icv Aβ42 infusion induced FLNA-α7nAChR and -TLR4 complexes											
2	Mice											
3	2-week C0105, 10 mg/kg b.i.d. on 7-day icv Aβ42 infusion											
4	Hippocampus						Prefrontal cortex					
5						FLNA						FLNA
6		OD	OD/FLNAOD	OD	OD/FLNAOD	OD		OD	OD/FLNAOD	OD	OD/FLNAOD	OD
7	Sham-veh	180	0.18	201	0.20	998	Sham-veh	249	0.35	121	0.17	712
8	Sham-veh	143	0.16	163	0.19	874	Sham-veh	309	0.29	204	0.19	1059
9	AVE	161.5	0.17	182.0	0.19		AVE	279.0	0.32	162.5	0.18	
10	SEM	9.24	0.01	9.49	0.00		SEM	14.99	0.04	20.74	0.01	
11												
12	sham-C105	269	0.27	223	0.22	994	sham-C105	233	0.26	124	0.14	897
13	sham-C105	243	0.30	152	0.19	817	sham-C105	286	0.27	196	0.18	1067
14	AVE	167.6	0.28	155.1	0.17		AVE	228.5	0.26	138.0	0.15	
15	SEM	32.32	0.02	26.78	0.03		SEM	38.18	0.01	23.73	0.03	
16												
17	ICV-veh	1319	1.33	1240	1.25	993	ICV-veh	1231	1.27	893	0.92	968
18	ICV-veh	1272	1.38	1013	1.10	924	ICV-veh	1355	1.36	932	0.94	994
19	ICV-veh	1073	1.15	916	0.98	932	ICV-veh	1347	1.41	1022	1.07	953
20	ICV-veh	1094	1.12	849	0.87	979	ICV-veh	1299	1.34	731	0.76	966
21	ICV-veh	1451	1.85	775	0.99	783	ICV-veh	1187	1.13	1026	0.97	1054
22	ICV-veh	1372	1.51	806	0.85	907	ICV-veh	1159	1.07	1012	0.94	1082
23	ICV-veh	1208	1.43	763	0.95	846	ICV-veh	1069	1.10	916	0.95	969
24	ICV-veh	1165	1.40	806	0.91	835	ICV-veh	1192	1.21	933	0.94	988
25	AVE	1244.3	1.40	896.0	0.99		AVE	1229.9	1.24	933.1	0.94	
26	SEM	47.20	0.08	60.16	0.05		SEM	34.91	0.05	34.18	0.03	
27												
28	ICV-C105	644	0.68	196	0.21	945	ICV-C105	528	0.55	253	0.26	964
29	ICV-C105	572	0.56	213	0.21	1015	ICV-C105	591	0.59	247	0.25	995
30	ICV-C105	328	0.34	175	0.18	979	ICV-C105	608	0.57	239	0.22	1066
31	ICV-C105	203	0.21	167	0.17	966	ICV-C105	559	0.52	288	0.27	1079
32	ICV-C105	459	0.59	194	0.25	772	ICV-C105	451	0.47	312	0.32	969
33	ICV-C105	379	0.50	183	0.24	764	ICV-C105	586	0.55	280	0.26	1073
34	ICV-C105	221	0.29	182	0.24	759	ICV-C105	561	0.51	206	0.19	1101
35	AVE	400.86	0.45	187.14	0.21		AVE	554.86	0.54	260.71	0.25	
36	SEM	63.35	0.07	5.75	0.01		SEM	19.89	0.02	13.31	0.02	
37			67.5%		78.3%				56.7%		73.0%	
38			9.9%		4.3%				1.7%		4.5%	
39												
40												

Furthermore, the method used to calculate the errors for these sham groups is incorrect. When calculating errors, the Standard Error of the Mean (SEM) is the Standard Deviation of the data (STDEV function in Excel) divided by the square root of N. It is not considered valid to calculate an SEM with N<3 samples. However, as shown here → the SEM has been calculated for 2 samples, by dividing the STDEV by 2.83, which is the square root of 8. Thus, errors have been calculated as if there were 8 samples, instead of 2.



fx =STDEV(E1:E8)/2.83

C	D	E	F
nt of the icv Aβ4		2 infusion in	duce
g/kg b.i.d. on		7-day icv A	β42 i
			FLN
FLNAOD	OD	OD/FLNAOD	OD
0.18	201	0.20	99
0.16	163	0.19	87
0.17	182.0	0.19	
0.01	9.49	E8)/2.83	
0.27	223	0.22	99

Lack of Original Images for Claimed Number of Replicates

Figure 6B contains data on insulin receptor (IR) signaling in control vs. Alzheimers brain samples treated with PTI-125. Accordingly, a .PPT file containing both the blots and graphs for this Figure is named "*IR signaling-C-AD 11 pairs quan-C0105.pp*t" (created 2011-05-16, 14:51, modified 2011-06-09, 13:55). This appears to be the earliest file with quantitative graphs used in the figure, so all western blot raw images used for the quantitation must have been completed by this date.

A search for blot images prior to 2011-06-09 reveals a collection of JPGs with "PM" (post-mortem) in their filenames, which contain the source images for Figure 6B. As seen elsewhere, the raw image and white-box image differ in each case by having "-1" appended to the end of the filename.

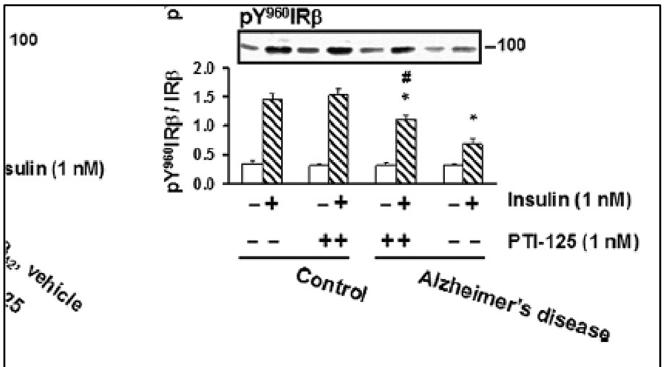
> *"PTI-PM-IR1-4.jpg"* (modified 2011-05-13, 14:38) is a raw image, and its corresponding white-box image *"PTI-PM-IR1-4-1.jpg"* (modified 2011-05-16, 21:31) appears to be the source for the IRβ blot (uppermost panel) in Figure 6B.
>
> *"pti-pm-pYIRb-IRS1R1-4.jpg"* (modified 2011-05-10, 17:48) is a raw image, and its corresponding white-box image *"pti-pm-pYIRb-IRS1R1-4-1.jpg"* (modified 2011-05-16, 21:17) appears to be the source for the IRS-1 and $pY^{1150/1151}$IRβ blots (2nd and 3rd from the top) in Figure 6B.
>
> "pti-pm-pYI9721-4.jpg" (modified 2011-05-10, 17:48) is a raw image, and its corresponding white-box image "pti-pm-pYI9721-4-1.jpg" (modified 2011-05-17, 10:28) appears to be the source for the pY^{960}IRβ blot (lower panel) in Figure 6B.

In addition to these 6 files, whose filenames contain "1-4", an additional 3 files were sourced from the same dates, with filenames containing "5-9". Each image named "1-4" contains 4 blots, while each image named "5-9" contains 5 blots, so it is reasonable to assume these files contain 9 separate experiments, each represented by a single blot.

The legend for Figure 6B (shown on the right) claims N=11, and there were 8 separate groups examined… control vs. Alzheimer's, +/- insulin, and +/- PTI-125. So, a minimum of 88 samples (11 x 8) would be required to provide the number of replicates claimed in the legend.

The .XLS spreadsheet embedded in the .PPT file for this figure indeed shows 11 separate values for each condition. However, each pair of raw images (tagged 1-4 and 5-9) does not contain a sufficient number of bands to account for the required number of 88 samples.

If the 88 samples were split across the 9 blots (1-4 & 5-9), each blot only has 10 lanes, and the blot images appear to show MW ladders in the first lane of each gel. So, there would be a maximum of 9 x 9 = 81 slots available, for the 88



d the $A\beta_{42}$-induced impairment in IR signaling in ICV $A\beta_{42}$-infused mice, as mea-
sociation with the signaling adaptor molecule IRS-1. $n = 7$ or 8. $^{*}p < 0.01$ vs sham,
ly, the deficit in IR signaling seen in postmortem AD and $A\beta_{42}$-treated control brain
$n = 11$. $^{*}p < 0.01$ vs vehicle-treated control group; $^{\#}p < 0.01$ vs vehicle-treated

samples to be loaded in. Alternatively, if all 88 samples were indeed run on these 9 blots, there would be no room for MW marker lanes on each gel, and yet the Figure shows MW markers alongside the blots? Ergo, the claimed number of samples (88) cannot possibly have been run on the 9 blots (1-4 & 5-9) for which primary images are available.

It cannot be ruled out that other .JPG images exist, containing the missing replicates, but this seems highly unlikely. The existing filenames suggest tags “1-4” and “5-9” refer to replicates, so there should be another 2 files containing replicates 10 and 11. There are no .JPG files with the same naming convention that contain the term “10-11”.

The earliest file appearing to contain data for Figure 6B is *“pti-pm-pYI9721-4.jpg”* (modified 2011-05-10, 17:48), and as discussed above the quantitation for the graphs took place on 2011-06-09, so all the blots should have been performed between these dates.

Only 35 .JPG files exist between May 10th and June 9th 2011. Nine of them are included in the analysis above, and a further 12 are tagged “ICV”, referring to the $A\beta_{42}$ infusion mouse model. The remainder are accounted for by other experiments, blotting for proteins not seen in Figure 6B. 3 are tagged “nNOS”, 5 tagged “nr”, 3 tagged “PLCg”, and 3 tagged “pYsrc”.

In other words, the available images do not contain sufficient data to account for the claim of N=11, and there do not appear to be any additional images that contain such data. Such images cannot exist after June 9th, because that is the cut-off date for quantitation to occur. If the images exist before May 10th, it would suggest that replicates #10 and #11 were run first, followed by 1-4 and 5-9. A search for the terms “C-AD” and “PM” in filenames also failed to find any other appropriate source images across a much earlier time frame extending back to January 2010. While it cannot be ruled out that such images exist elsewhere, this seems highly unlikely.

Summary

A terminal digit analysis on the source data for the 2012 *J. Neurosci.* paper indicates there is a 1 in 100 million probability that the numerical data used to construct the graphs in the paper came about from a random stochastic process that would be consistent with true biological experimentation. Data underlying several individual figures is similarly problematic.

In the case of Figure 1, the number of biological replicates (7–8) claimed in the figure legend is not supported by the underlying source data which shows a much smaller number (2). The methods used to calculate error bars for the graphs in this figure are also inappropriate.

In the case of Figure 6, the underlying source data indeed show the appropriate number of replicates, matching the claim in the Figure legend (11). However, the .JPG source images that would have been used for generating this data (quantitation by densitometry) do not contain the required number of bands or gels. A maximum of 9 experimental replicates appears possible, and file naming conventions plus date-restrictions preclude the possibility that the *missing* data are likely to be found elsewhere.

The preponderance of evidence therefore suggests several graphs in the 2012 *J. Neurosci.* paper contain numerical data that has been fabricated. The generation of such data by human hand rather than stochastic experimentation appears to have introduced non-natural terminal digit distributions, with a preponderance for the number 9 as a terminal digit.

Paul Brooke

Paul S. Brookes, PhD.